

ON THE ARBITRARILY LONG-TERM STABILITY OF CONSERVATIVE METHODS*

ANDY T. S. WAN[†] AND JEAN-CHRISTOPHE NAVE[‡]

Abstract. We show the arbitrarily long-term stability of conservative methods for autonomous ODEs. Given a system of autonomous ODEs with conserved quantities, if the preimage of the conserved quantities possesses a bounded locally finite neighborhood, then the global error of any conservative method is bounded for all time, provided the uniform time step is taken sufficiently small. On finite precision machines, the global error still remains bounded until some arbitrarily large time determined by machine precision and tolerance. The main result is proved using elementary topological properties for discretized conserved quantities which are equicontinuous. In particular, stability is also shown using an averaging identity when the discretized conserved quantities do not explicitly depend on time steps. In addition, we derive a sufficient condition for constructing conservative methods using the multiplier method and give explicit formulas for first order conservative methods in the case of polynomial conserved quantities. Numerical results are shown to illustrate the main stability result.

Key words. conservative, conservative methods, long-term stability, finite difference, multiplier method, autonomous system

AMS subject classifications. 65L05 65L12, 65L20, 65L70, 65P10, 65Z05

1. Introduction. In recent years, there has been vast renewed interests in structure-preserving discretizations; that is numerical methods which preserve underlying structures of differential equations at the discrete level [19, 22, 15, 17, 14, 4, 1, 8]. One primary motivation for these discretizations is, for some class of problems, the ability to preserve certain features inherent to the continuous problem is a determining factor for acceptance of numerical results. For instance, for ODEs with a Hamiltonian structure, preservation of phase space volume is a desirable feature for the discrete flow of symplectic methods [14]. For systems arising from variational formulation, the variational principle is preserved by variational integrators via extremizing the action integral over a finite dimensional discrete space [19]. Beyond this primary motivation, structure-preserving discretizations can also possess additional stability and long-term properties. For example, symplectic methods have been shown to possess favorable long-term properties, such as near conservation of energy over an exponentially long time [2]. Moreover, for completely integrable Hamiltonian systems, symplectic methods nearly conserve all first integrals depending only on action variables and have at most linear growth in the global error over an exponentially long time [5, 6, 14].

In the present work, we focus on the question pertaining to stability properties on the class of *conservative methods*; specifically discretizations which exactly preserve conserved quantities at the discrete level. Conservative methods for ODEs and PDEs have a long history in numerical analysis [9, 16, 24, 18, 13, 23, 12, 10, 11, 7]. Traditionally, conservative methods have been proposed for various types of equations and special forms of conserved quantities. Unfortunately, it is perhaps due to the specialization of these methods that fundamental stability properties may have been overlooked. Recently, a general conservative method, called the multiplier method

*This work was supported by the NSERC Discovery program and the Centre de Recherches Mathématiques.

[†]Department of Mathematics and Statistics, McGill University, Montréal, QC, H3A 0B9, Canada (andy.wan@mcgill.ca).

[‡]Department of Mathematics and Statistics, McGill University, Montréal, QC, H3A 0B9, Canada (jcnav@math.mcgill.ca).

[25], has been proposed for discretizing ODEs and PDEs so that their underlying conserved quantities can be exactly preserved at the discrete level. Motivated by the general applicability of the multiplier method, it is important, in our view, to provide a general stability result for conservative methods; specifically for the case of ODEs. On the outset, this may seem like an impossible task as most discretizations constructed by the multiplier method do not possess an a priori common structure and are often nonlinear in nature. Fortunately, these difficulties can be resolved when one takes the point of view that *long-term stability* is intimately connected with the topology of the conserved quantities. Specifically for ODEs, under some appropriate conditions, we will show using basic topological arguments that the global error of a conservative method is *bounded for all time*.

This paper is organized as follows. In Section 2, we review basic properties of conservative methods. It is shown that equicontinuity of discretized conserved quantities plays a central role in many estimates used in subsequent sections. In Section 3, we review some elementary topology results relevant to our current discussion and show a key separation theorem which form the basis for the main stability result. Moreover, we discuss the practical limitation of the main result in finite precision arithmetic. In Section 4, we derive a sufficient condition for conservative method using the multiplier method and provide an explicit formula in the case of polynomial conserved quantities. Finally in Section 5, we verify the main stability result numerically and make comparison with traditional and symplectic methods.

2. Preliminaries.

2.1. Definitions and notations. Let $n \in \mathbb{N}$ and U be an open subset of \mathbb{R}^n . Suppose $\mathbf{f} : U \rightarrow \mathbb{R}^n$ is locally Lipschitz continuous. Then by Picard's theorem, for any $\mathbf{x}_0 \in U$, there exists an open interval $I = (-T, T)$ such that the autonomous ODE,

$$(1) \quad \begin{aligned} \mathbf{F}[\mathbf{x}]_t &:= \dot{\mathbf{x}}(t) - \mathbf{f}(\mathbf{x}(t)) = \mathbf{0}, \\ \mathbf{x}(0) &= \mathbf{x}_0, \end{aligned}$$

has an unique solution $\mathbf{x} \in C^1(I; U)$. For brevity, we used the notation $[\mathbf{x}]_t$ denoting dependence on $t, \mathbf{x}(t)$ and higher derivatives of $\mathbf{x}(t)$.

For $1 \leq m \leq n$. We assume the ODE (1) has m conserved quantities; that is there exists a continuous vector-valued function $\boldsymbol{\psi} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that for the constant $\mathbf{c} = \boldsymbol{\psi}(\mathbf{x}_0)$, the unique solution $\mathbf{x} \in C^1(I; U)$ satisfies for $t \in I$,

$$(2) \quad \boldsymbol{\psi}(\mathbf{x}(t)) = \mathbf{c}.$$

As to be discussed later, the preimage of \mathbf{c} , denoted as $\boldsymbol{\psi}^{-1}(\{\mathbf{c}\})$, can be written as,

$$\boldsymbol{\psi}^{-1}(\{\mathbf{c}\}) = \bigcup_{j \in J} X_j,$$

for some countable¹ index set J with each X_j as a connected component of $\boldsymbol{\psi}^{-1}(\{\mathbf{c}\})$. Denote the connected component containing \mathbf{x}_0 as X_0 . If X_0 is compact, then by standard theory of ODEs, the local solution \mathbf{x} can be extended to a global solution for all $t \in \mathbb{R}$.

¹The number of connected components of any subsets in \mathbb{R}^n is at most countable; See [20] or Section 3.

THEOREM 1. *Suppose the connected component X_0 containing \mathbf{x}_0 is compact, then the unique solution $\mathbf{x} \in C^1(I; U)$ to (1) can be extended for all $t \in \mathbb{R}$ and so $\mathbf{x} \in C^1(\mathbb{R}; X_0)$.*

Let $\tau > 0$ be a time step size and consider the set of uniform time steps of $\{t_k = k\tau : k \in \mathbb{N}\}$. For a given $\mu \in \mathbb{N}$, we shall consider general μ -step discretizations or μ -step methods (we use both terms interchangeably). In particular, let $\mathbf{F}^\tau : \mathbb{R}^{n(\mu+1)} \rightarrow \mathbb{R}^n$ be a continuous vector-valued function depending on τ . Then the μ -step discretization is given by,

$$(3) \quad \mathbf{F}^\tau \{\mathbf{x}^\tau\}_k := \mathbf{F}^\tau(\mathbf{x}_{k+1}, \mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1}) = 0,$$

Similar to the continuous case, we employ the notation $\{\mathbf{x}^\tau\}_k$ to denote dependence on the successive approximation \mathbf{x}_k at different time steps. For technical reasons, we shall consider discretizations which have the *local contraction property*.

DEFINITION 2. *A μ -step discretization (3) has a local contraction property if for any compact subset $K \subset \mathbb{R}^n$ with $\mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1} \in K$ and for any $r > 0$, there exists $\tau_c > 0$ depending on r and K such that for $0 < \tau < \tau_c$, the discretization (3) have an unique solution $\mathbf{x}_{k+1} \in \bigcup_{i=0}^{\mu-1} B_r(\mathbf{x}_{k-i})$.*

For example, local contraction property of a μ -step discretization can typically be shown via a fixed point argument for implicit methods.

DEFINITION 3. *A μ -step discretization \mathbf{F}^τ is consistent of order $p > 0$ to \mathbf{F} if for $\mathbf{x} \in C^{p+1}(I; U)$ and each time step t_k , there exists a positive constant C_F independent of τ such that,*

$$\|\mathbf{F}[\mathbf{x}]_{t_k} - \mathbf{F}^\tau(\mathbf{x}(t_{k+1}), \mathbf{x}(t_k), \dots, \mathbf{x}(t_{k-\mu+1}))\| \leq C_F \|\mathbf{x}\|_{C^{p+1}(\tilde{I}_k)} \tau^p,$$

where $\tilde{I}_k = [t_{k-\mu+1}, t_{k+1}]$ and $\|\mathbf{x}\|_{C^{p+1}(\tilde{I}_k)} = \max_{0 \leq i \leq p+1} \left\| \frac{d^i \mathbf{x}}{dt^i} \right\|_{L^\infty(\tilde{I}_k)}$.

In practice, C_F arise from Taylor expansion with remainder terms. Similarly, let $\psi^\tau : \mathbb{R}^{n\mu} \rightarrow \mathbb{R}^m$ be a continuous vector-valued function depending on τ .

DEFINITION 4. *The discrete conserved quantities ψ^τ is consistent to ψ of order p if for $\mathbf{x} \in C^p(I; U)$ and each time step t_k , there exists a positive constant C_ψ independent of τ such that,*

$$\|\psi(\mathbf{x}(t_k)) - \psi^\tau(\mathbf{x}(t_k), \dots, \mathbf{x}(t_{k-\mu+1}))\| \leq C_\psi \|\mathbf{x}\|_{C^p(I_k)} \tau^p,$$

where $I_k = [t_{k-\mu+1}, t_k]$ and $\|\mathbf{x}\|_{C^p(I_k)} = \max_{0 \leq i \leq p} \left\| \frac{d^i \mathbf{x}}{dt^i} \right\|_{L^\infty(I_k)}$

2.2. Equicontinuity and averaging identity. If a family of discretized conserved quantities $\{\psi^\tau\}_{0 < \tau < \tau_0}$ is equicontinuous for some $\tau_0 > 0$, one immediate consequence is the following important relation between ψ and ψ^τ .

LEMMA 5. *Suppose $\psi^\tau : \mathbb{R}^{n\mu} \rightarrow \mathbb{R}^m$ is consistent of order p to ψ and for some $\tau_0 > 0$, the family of functions $\{\psi^\tau\}_{0 < \tau < \tau_0}$ is equicontinuous on $U \times \dots \times U \subset \mathbb{R}^{n\mu}$ for some open subset $U \subset \mathbb{R}^n$. Then for any $\mathbf{y} \in U$,*

$$\psi(\mathbf{y}) = \lim_{\tau \rightarrow 0} \psi^\tau(\mathbf{y}, \dots, \mathbf{y}).$$

Proof. For any $\epsilon > 0$ and $\mathbf{y} \in U$, pick a function $\mathbf{x}(t) \in C^{(p)}([t_{k-\mu+1}, t_k])$ with $\mathbf{x}(t_k) = \mathbf{y}$. For fixed t_k , by continuity of $\mathbf{x}(t)$, there exists a positive constant τ_1 such that if $0 < \tau < \tau_1$, then $\mathbf{x}(t_{k-i}) \in U$ for $i = 0, \dots, \mu - 1$. Thus, by equicontinuity and since all norms are equivalent on $\mathbb{R}^{n\mu}$, there exists a positive constant δ depending only on ϵ such that if $\max_{0 \leq i \leq \mu-1} \|\mathbf{y} - \mathbf{x}(t_{k-i})\| < \delta$,

$$\|\psi^\tau(\mathbf{x}(t_k), \mathbf{x}(t_{k-1}), \dots, \mathbf{x}(t_{k-\mu+1})) - \psi^\tau(\mathbf{y}, \mathbf{y}, \dots, \mathbf{y})\| \leq \frac{\epsilon}{2},$$

for all $0 < \tau < \min\{\tau_0, \tau_1\}$. Moreover, by continuity of \mathbf{x} again, there exists some positive constant τ_2 such that if $0 < \tau < \tau_2$, $\|\mathbf{y} - \mathbf{x}(t_{k-i})\| < \delta$ for all $i = 0, \dots, \mu - 1$. Combining together with consistency, this implies for $0 < \tau <$

$$\min \left\{ \tau_0, \tau_1, \tau_2, \left(\frac{\epsilon}{2C_\psi \|\mathbf{x}\|_{C^p(I_k)}} \right)^{\frac{1}{p}} \right\},$$

$$\begin{aligned} \|\psi(\mathbf{y}) - \psi^\tau(\mathbf{y}, \dots, \mathbf{y})\| &\leq \|\psi(\mathbf{x}(t_k)) - \psi^\tau(\mathbf{x}(t_k), \mathbf{x}(t_{k-1}), \dots, \mathbf{x}(t_{k-\mu+1}))\| \\ &\quad + \|\psi^\tau(\mathbf{x}(t_k), \mathbf{x}(t_{k-1}), \dots, \mathbf{x}(t_{k-\mu+1})) - \psi^\tau(\mathbf{y}, \mathbf{y}, \dots, \mathbf{y})\| \\ &\leq C_\psi \|\mathbf{x}\|_{C^p(I_k)} \tau^p + \frac{\epsilon}{2} < \epsilon. \end{aligned}$$

In other words, the limit $\psi^\tau(\mathbf{y}, \dots, \mathbf{y})$ as $\tau \rightarrow 0$ exists and is equal to $\psi(\mathbf{y})$. \square

If ψ^τ does not depend on τ explicitly, then Theorem 5 follows immediately, as ψ^τ is trivially equicontinuous on U . In fact, the following remarkably simple identity holds.

COROLLARY 6 (Averaging Identity). *Suppose $\psi^\tau : \mathbb{R}^{n\mu} \rightarrow \mathbb{R}^m$ is consistent of order p to ψ and assume ψ^τ does not depend on τ explicitly. Then for any $\mathbf{y} \in \mathbb{R}^n$,*

$$\psi(\mathbf{y}) = \psi^\tau(\mathbf{y}, \dots, \mathbf{y}).$$

Proof. Let $\mathbf{y} \in \mathbb{R}^n$ and t_k be fixed and let $\mathbf{x}(t) \in C^{(p)}([t_{k-\mu+1}, t_k])$ with $\mathbf{x}(t_k) = \mathbf{y}$. By continuity of \mathbf{x} and for fixed t_k , $\lim_{\tau \rightarrow 0} \mathbf{x}(t_{k-i}) = \mathbf{y}$ for all $i = 0, \dots, \mu - 1$. Since ψ^τ does not depend on τ explicitly, then by consistency and continuity of ψ^τ ,

$$\begin{aligned} \|\psi(\mathbf{y}) - \psi^\tau(\mathbf{y}, \mathbf{y}, \dots, \mathbf{y})\| &= \lim_{\tau \rightarrow 0} \|\psi(\mathbf{y}) - \psi^\tau(\mathbf{y}, \mathbf{y}, \dots, \mathbf{y})\| \\ &\leq \lim_{\tau \rightarrow 0} \|\psi(\mathbf{y}) - \psi^\tau(\mathbf{y}, \mathbf{x}(t_{k-1}), \dots, \mathbf{x}(t_{k-\mu+1}))\| \\ &\quad + \lim_{\tau \rightarrow 0} \|\psi^\tau(\mathbf{y}, \mathbf{x}(t_{k-1}), \dots, \mathbf{x}(t_{k-\mu+1})) - \psi^\tau(\mathbf{y}, \mathbf{y}, \dots, \mathbf{y})\| \\ &\leq \underbrace{\lim_{\tau \rightarrow 0} C_\psi \|\mathbf{x}\|_{C^p(I_k)} \tau^p}_{=0} \\ &\quad + \underbrace{\left\| \psi^\tau(\mathbf{y}, \lim_{\tau \rightarrow 0} \mathbf{x}(t_{k-1}), \dots, \lim_{\tau \rightarrow 0} \mathbf{x}(t_{k-\mu+1})) - \psi^\tau(\mathbf{y}, \mathbf{y}, \dots, \mathbf{y}) \right\|}_{=0} \end{aligned} \quad \square$$

To the best knowledge of the authors, we have not seen this remarkably simple identity relating ψ and ψ^τ appeared in the previous literature. The term *averaging identity* originates from applications where such ψ^τ can typically be interpreted as a kind of (nonlinear) average of ψ among different time steps t_k .

2.3. Conservative discretization.

DEFINITION 7. *The discretization (3) is called conservative if*

$$\psi^\tau\{\mathbf{x}^\tau\}_{k+1} = \psi^\tau\{\mathbf{x}^\tau\}_k, \text{ for } k \in \{\mu-1, \mu, \dots\}.$$

For a conservative discretization, it follows by induction that for $k \in \{\mu-1, \mu, \dots\}$,

$$(4) \quad \psi^\tau\{\mathbf{x}^\tau\}_k = \psi^\tau(\mathbf{x}_k, \dots, \mathbf{x}_{k-\mu}) = \psi^\tau(\mathbf{x}_{\mu-1}, \dots, \mathbf{x}_0) =: \mathbf{c}^\tau.$$

In the case of 1-step methods with ψ^τ not explicitly depending on τ , then the averaging identity of Lemma 5 implies $\psi^\tau(\mathbf{y}) = \psi(\mathbf{y})$ and so $\mathbf{c}^\tau = \mathbf{c}$. For general μ -step methods, $\mu-1$ initial values must be specified in order to proceed. This initialization step is usually handled by using one-step methods of sufficient order, such as Runge-Kutta methods. However, since traditional 1-step methods are generally not conservative, there will be a corresponding error in the constant $\mathbf{c}^\tau = \psi^\tau(\mathbf{x}_{\mu-1}, \dots, \mathbf{x}_0)$. Fortunately, as we show in Section 3, this initialization error does not pose a problem for the long-term stability result, as long as we can choose the error in $\|\mathbf{c} - \mathbf{c}^\tau\|$ to be arbitrarily small. In particular, we need the following result in subsequent section.

LEMMA 8. *Let $\psi^\tau : \mathbb{R}^{n\mu} \rightarrow \mathbb{R}^n$ be consistent of order p to ψ . Suppose the μ initial values are p -th order accurate; that is for the unique solution $\mathbf{x} \in C^1(I; U) \cap C^p([0, t_{\mu-1}]; U)$ to the ODE (1), the μ initial values $\{\mathbf{x}_k\}_{k=0}^{\mu-1}$ satisfies for some positive constants C, τ_μ independent of τ such that if $0 < \tau < \tau_\mu$,*

$$(5) \quad \max_{0 \leq k \leq \mu-1} \|\mathbf{x}(t_k) - \mathbf{x}_k\| \leq C\tau^p.$$

Also assume for some $\tau_0 > 0$, the family of functions $\{\psi^\tau\}_{0 < \tau < \tau_0}$ is equicontinuous on $U \times \dots \times U \subset \mathbb{R}^{n\mu}$. Then for $\mathbf{c} = \psi(\mathbf{x}_0)$,

$$\lim_{\tau \rightarrow 0} \|\mathbf{c} - \mathbf{c}^\tau\| = 0.$$

Proof. The proof is similar to Lemma 5. Let $\epsilon > 0$ and \mathbf{x} be the exact solution to the ODE (1) and $\{\mathbf{x}_k\}_{k=0}^{\mu-1}$ be the given initial values. From (5) and that $\mathbf{x}(t_k) \in U$ for all $0 \leq k \leq \mu-1$, it follows that for some positive constant τ_1 , $\mathbf{x}_k \in U$ for all $0 \leq k \leq \mu-1$ and $0 < \tau < \tau_1$. By equicontinuity, there exists a $\delta > 0$ depending only on ϵ such that if $\max_{0 \leq k \leq \mu-1} \|\mathbf{x}(t_k) - \mathbf{x}_k\| < \delta$, then

$$\|\psi^\tau(\mathbf{x}(t_{\mu-1}), \dots, \mathbf{x}(t_0)) - \psi^\tau(\mathbf{x}_{\mu-1}, \dots, \mathbf{x}_0)\| < \frac{\epsilon}{2},$$

for all $0 < \tau < \min\{\tau_0, \tau_1\}$. Indeed, $\max_{0 \leq k \leq \mu-1} \|\mathbf{x}(t_k) - \mathbf{x}_k\| < \delta$ is fulfilled by hypothesis (5) if $0 < \tau < \tau_2$ for some positive constant τ_2 . Since $\mathbf{c} = \psi(\mathbf{x}(t_{\mu-1}))$ by (2) and $\mathbf{c}^\tau = \psi^\tau(\mathbf{x}_{\mu-1}, \dots, \mathbf{x}_0)$ by (4), it follows from consistency that for sufficiently small τ ,

$$\begin{aligned} \|\mathbf{c} - \mathbf{c}^\tau\| &\leq \|\psi(\mathbf{x}(t_{\mu-1})) - \psi^\tau(\mathbf{x}(t_{\mu-1}), \dots, \mathbf{x}(t_0))\| \\ &\quad + \|\psi^\tau(\mathbf{x}(t_{\mu-1}), \dots, \mathbf{x}(t_0)) - \psi^\tau(\mathbf{x}_{\mu-1}, \dots, \mathbf{x}_0)\| \\ &\leq C_\psi \|\mathbf{x}\|_{C^p(I_k)} \tau^p + \frac{\epsilon}{2} < \epsilon. \end{aligned} \quad \square$$

3. Main results. We now discuss a long-term stability result for conservative methods. Although in application, we have in mind the underlying space is $X = \mathbb{R}^n$, which is sufficient for ODEs in finite dimensions. In anticipation for subsequent work on evolution PDEs, X can be some function space where the PDEs are viewed as ODEs over infinite dimensional spaces. Since the main ideas are mostly based on topological properties, we will state the main theorem in a general setting and restricting X to a metric space when necessary. First, we review some elementary results from topology relevant to our discussion. See [20] for more details.

THEOREM 9. *Let $A \subset X$ be a nonempty subset of a locally connected, second-countable topological space X . Then $A = \bigcup_{j \in J} A_j$ for some countable indexed set J , where the collection of A_j are connected components of A with each A_j being nonempty, closed in A and disjoint from each other.*

Theorem 9 implies immediately the following:

LEMMA 10. *Let X and Y be topological spaces with X locally connected and second-countable and Y Hausdorff. Suppose $\psi : X \rightarrow Y$ is a continuous function. For any $\mathbf{c} \in Y$ with a nonempty preimage $\psi^{-1}(\{\mathbf{c}\})$, there is some countable indexed set J such that,*

$$\psi^{-1}(\{\mathbf{c}\}) = \bigcup_{j \in J} X_j,$$

where each X_j is a nonempty, closed subset in $\psi^{-1}(\{\mathbf{c}\})$ and disjoint from each other.

Similarly, we will be interested in working with preimage of neighborhoods in metric spaces. Specifically, for a metric space Y with a metric $d_Y(\cdot, \cdot)$, we denote the open neighborhood $B_\epsilon(\mathbf{c}) = \{\mathbf{y} \in Y : d_Y(\mathbf{y}, \mathbf{c}) < \epsilon\}$.

LEMMA 11. *Let X be a locally connected, second-countable topological space and Y be a metric space. Suppose $\psi : X \rightarrow Y$ is a continuous function. For any $\mathbf{c} \in Y$ and any $\epsilon > 0$ with a nonempty preimage $\psi^{-1}(\overline{B_\epsilon(\mathbf{c})})$, there is some countable indexed set J^ϵ such that,*

$$\psi^{-1}(\overline{B_\epsilon(\mathbf{c})}) = \bigcup_{j \in J^\epsilon} X_j^\epsilon,$$

where each X_j^ϵ is a nonempty, closed subset in $\psi^{-1}(\overline{B_\epsilon(\mathbf{c})})$ and disjoint from each other.

LEMMA 12. *Let X be a topological space. If $A \subset X$ is closed in X and $B \subset A$ is closed in A , then B is closed in X .*

LEMMA 13. *If the hypotheses of Lemma 10 are satisfied, then each X_j is a closed subset of X . Moreover, if Y is a metric space, then each X_j^ϵ is closed in X for any $\epsilon > 0$.*

Proof. By continuity of ψ and $\{\mathbf{c}\}$ is closed in Y (since Y is Hausdorff), $\psi^{-1}(\{\mathbf{c}\})$ is closed in X . Since X_j is closed in $\psi^{-1}(\{\mathbf{c}\})$, then Lemma 12 implies X_j is closed in X . The proof proceeds similarly for X_j^ϵ . \square

3.1. Locally finite neighborhood. For the main stability result, we wish to include cases where the connected components can be bounded or unbounded. Moreover, we also wish to handle the possibility of countably infinitely many connected components. However, it turns out the main stability result hinges on whether certain

connected components can be separated by open neighborhoods. In particular, there are situations which can arise we wish to exclude, as the following example illustrate.

EXAMPLE 1. Let $X = \mathbb{R} = Y$ and consider the smooth function:

$$\psi(x) = \begin{cases} \exp\left(-\frac{1}{x^2}\right) \sin\left(\frac{1}{x^2}\right), & x \neq 0 \\ 0, & x = 0 \end{cases}$$

The preimage $\psi^{-1}(\{0\})$ has the connected components $\{0\} \cup \bigcup_{k \in \mathbb{N}} \left\{\pm \frac{1}{\sqrt{k\pi}}\right\}$. But for $\epsilon > 0$, the neighborhood $(-\epsilon, \epsilon)$ around $X_0 = \{0\}$ intersects infinite many connected components $\left\{\pm \frac{1}{\sqrt{k\pi}}\right\}$ for all $k > \frac{1}{\epsilon^2\pi}$.

The preceding example shows that X_0 cannot be separated by open neighborhoods if it has “too many” neighboring connected components. This leads to the following definition.

DEFINITION 14. Let $\{U_\beta\}_{\beta \in J}$ be a collection of closed subsets of a topological space X with an index set J . For a fixed $\alpha \in J$, U_α is said to have a **locally finite neighborhood (LFN)** V if V is an open subset of X such that:

- $U_\alpha \subset V$
- $U_\beta \cap V = \emptyset$ for all but finitely many $\beta \in J$

Furthermore, if X is a metric space, we say that V is a **bounded LFN** of U_α if V is also bounded.

For a normal topological space X , an equivalent definition of a LFN is that U_α and $\bigcup_{\beta \neq \alpha} U_\beta$ can be separated by open neighborhoods.

THEOREM 15. Let X be a normal topological space and $\{U_\beta\}_{\beta \in J}$ be a collection of closed subsets in X . Then U_α has a LFN V if and only if there exists disjoint open subsets A, B in X such that $U_\alpha \subset A \subset V$ and $\bigcup_{\alpha \neq \beta \in J} U_\beta \subset B$.

Proof. It suffices to prove only the forward implication, since if there exists such disjoint open subsets A, B , then U_α has a LFN A . Suppose U_α has a LFN V so that $U_\alpha \subset V$ and at most a finite collection $\{U_\beta\}_{\beta \in J'}$ with $\alpha \notin J'$ such that $U_\beta \cap V \neq \emptyset$ for all $\beta \in J'$. Since J' is finite, $\bigcup_{\beta \in J'} (U_\beta \cap \bar{V})$ is closed in X . Since U_α is closed in a normal topological space X , there exists disjoint open subsets A', B' in X such that $U_\alpha \subset A'$ and $\bigcup_{\beta \in J'} (U_\beta \cap \bar{V}) \subset B'$. Let $A = A' \cap V$ and $B = B' \cup (X - \bar{V})$. Then clearly both A, B are disjoint and open in X with $U_\alpha \subset A \subset V$. Thus, the result follows since,

$$\bigcup_{\alpha \neq \beta \in J} U_\beta = \underbrace{\left(\bigcup_{\beta \in J'} (U_\beta \cap \bar{V}) \right)}_{\subset B'} \cup \underbrace{\left(\bigcup_{\beta \in J'} (U_\beta \cap (X - \bar{V})) \right)}_{\subset X - \bar{V}} \cup \underbrace{\left(\bigcup_{\substack{\beta \in J \\ \beta \notin J'}} U_\beta \right)}_{\subset X - \bar{V}} \subset B. \quad \square$$

COROLLARY 16. Let X be a metric space and Y be a Hausdorff topological space. Suppose $\psi : X \rightarrow Y$ is continuous function with a nonempty preimage $\psi^{-1}(\{\mathbf{c}\}) = \bigcup_{j \in J} X_j$ for some countable index set J . Then X_0 has a LFN V if and only if there exists disjoint open subsets A, B in X such that $X_0 \subset A \subset V$ and $\bigcup_{0 \neq j \in J} X_j \subset B$. Thus, if X_0 has a bounded LFN V , then A is also bounded.

Proof. Since any metric space X is locally connected and second-countable, X_j is closed in X by Lemma 13 for all $j \in J$. As X is also normal, applying Theorem 15 for the collection of closed subsets $\{X_j\}_{j \in J}$ implies the result. \square

Furthermore, we will need the following two lemmas regarding compact subsets.

LEMMA 17. *Let X be a topological space and let $\{A_n\}_{n \in \mathbb{N}}$ be a decreasing nested sequence of nonempty compact subsets in X . For any open subset U in X such that $\bigcap_{n \in \mathbb{N}} A_n \subset U$, there exists a positive integer N such that $A_n \subset U$ for all $n \geq N$.*

Proof. By Cantor's intersection theorem, there exists $\mathbf{x} \in \bigcap_{n \in \mathbb{N}} A_n$. Now assume the contrary, then there exists a sequence $n_i \rightarrow \infty$ such that $A_{n_i} \cap (X - U) \neq \emptyset$. Clearly, $\bigcap_{n \in \mathbb{N}} A_n \subset \bigcap_{i \in \mathbb{N}} A_{n_i}$. Moreover, for all $n \in \mathbb{N}$, there exists $n_i \geq n$ so that $A_{n_i} \subset A_n$, since A_n are decreasing and $n_i \rightarrow \infty$. This implies $\bigcap_{i \in \mathbb{N}} A_{n_i} \subset \bigcap_{n \in \mathbb{N}} A_n$. It follows that $\mathbf{x} \in \bigcap_{n \in \mathbb{N}} A_n = \bigcap_{i \in \mathbb{N}} A_{n_i} \subset X - U$ which contradicts $\mathbf{x} \in \bigcap_{n \in \mathbb{N}} A_n \subset U$. \square

LEMMA 18. *Let X be a metric space and A, B be subsets of X with A compact in X and $A \cap \overline{B} = \emptyset$. Then $d_X(A, B) > 0$.*

Proof. Suppose not, then there is a sequence $a_n \in A$ such that $d_X(a_n, B) \rightarrow 0$. Since A is compact, there is a convergent subsequence $a_{n_i} \rightarrow a \in A$. Thus, $d_X(a, B) = \lim_{i \rightarrow \infty} d_X(a_{n_i}, B) = 0$ or in other words $a \in \overline{B}$ which contradicts that $A \cap \overline{B} = \emptyset$. \square

Finally, we show a key separation theorem for establishing the main stability theorem for conservative methods. Note that for any $\epsilon > 0$, since $X_0 \subset \psi^{-1}(\overline{B_\epsilon(\mathbf{c})})$, $X_0 \subset X_j^\epsilon$ for some index j in J^ϵ . By rearranging J^ϵ if necessary, we can denote X_0^ϵ to be the unique connected component containing X_0 for all $\epsilon > 0$.

THEOREM 19. *Let X, Y be metric spaces and let $\psi : X \rightarrow Y$ be a continuous function with a nonempty preimage $\psi^{-1}(\{\mathbf{c}\}) = \bigcup_{j \in J} X_j$ for some countable index set J . For each $\epsilon > 0$, denote the nonempty preimage $\psi^{-1}(\overline{B_\epsilon(\mathbf{c})}) = \bigcup_{j \in J^\epsilon} X_j^\epsilon$ for some countable index set J^ϵ . Suppose X_0 has a bounded LFN V with \overline{V} compact, then there exists $\epsilon_0 > 0$ such that if $0 < \epsilon < \epsilon_0$, X_0^ϵ is compact and is separated from $\bigcup_{0 \neq j \in J^\epsilon} X_j^\epsilon$.*

Proof. Let $A \subset V$ and B be such disjoint open sets from Corollary 16. Now suppose for all $\epsilon > 0$, there exists $\mathbf{x} \in \psi^{-1}(\overline{B_\epsilon(\mathbf{c})}) \cap (X - (A \cup B))$. Then for all $\epsilon > 0$, $\mathbf{x} \in \psi^{-1}(\overline{B_\epsilon(\mathbf{c})})$, or equivalently $\psi(\mathbf{x}) = \mathbf{c}$. This implies $\mathbf{x} \in X_j \subset A \cup B$ for some $j \in J$, which contradicts $\mathbf{x} \in X - (A \cup B)$. It follows that there exists $\epsilon' > 0$ so that if $0 < \epsilon \leq \epsilon'$,

$$\bigcup_{j \in J^\epsilon} X_j^\epsilon = \psi^{-1}(\overline{B_\epsilon(\mathbf{c})}) \subset A \cup B.$$

Since A, B are disjoint and $X_j^{\epsilon'}$ is connected for any $j \in J^{\epsilon'}$, either $X_j^{\epsilon'} \subset A$ or $X_j^{\epsilon'} \subset B$. In the case when $j = 0$, then $X_0^{\epsilon'} \subset A$, since otherwise $X_0 \subset X_0^{\epsilon'} \subset B$ which contradicts $X_0 \cap B = \emptyset$. Moreover, $X_0^{\epsilon'}$ is compact, since $X_0^{\epsilon'}$ is closed by Lemma 13 and $X_0^{\epsilon'} \subset A \subset V \subset \overline{V}$ with \overline{V} compact. Similarly for $0 \neq j \in J^{\epsilon'}$, in the case if $X_j^{\epsilon'} \subset A$, then $X_j^{\epsilon'} \subset A - X_0^{\epsilon'}$, since $X_j^{\epsilon'}$ and $X_0^{\epsilon'}$ are disjoint if $j \neq 0$. Thus, for any $0 < \epsilon \leq \epsilon'$,

$$(6) \quad \bigcup_{0 \neq j \in J^\epsilon} X_j^\epsilon \subset \bigcup_{0 \neq j \in J^{\epsilon'}} X_j^{\epsilon'} \subset (A - X_0^{\epsilon'}) \cup B$$

Now define the following two disjoint open subsets,

$$A' := \text{int}(X_0^{\epsilon'}), \quad B' := (A - X_0^{\epsilon'}) \bigcup B.$$

For the moment, assume the following claim is true:

CLAIM 20. *For some $\epsilon_0 \leq \epsilon'$, X_0^ϵ is compact and $X_0^\epsilon \subset A'$ for all $0 < \epsilon < \epsilon_0$.*

Thus, combining (6) and Claim 20 implies the theorem. It remains to show Claim 20 for which we proceed in two main steps. First, we show that,

$$(7) \quad X_0 \subset A'.$$

Indeed, let $\mathbf{x} \in X_0 \subset X_0^{\epsilon'}$, then by continuity of ψ , there exists $r > 0$ so that,

$$B_r(\mathbf{x}) \subset \psi^{-1}(B_{\epsilon'}(\mathbf{c})) \subset \bigcup_{j \in J^{\epsilon'}} X_j^{\epsilon'}$$

Since $B_r(\mathbf{x})$ is connected and $\{X_j^{\epsilon'}\}_{j \in J^{\epsilon'}}$ are connected components, $B_r(\mathbf{x}) \subset X_j^{\epsilon'}$ for some $j \in J^{\epsilon'}$. Supposing $j \neq 0$ implies the contradiction that $\mathbf{x} \in X_j^{\epsilon'} \cap X_0^{\epsilon'} = \emptyset$. So $B_r(\mathbf{x}) \subset X_0^{\epsilon'}$ or in other words \mathbf{x} is an interior point of $X_0^{\epsilon'}$ which implies (7).

Secondly, we show there exists $\epsilon_0 \leq \epsilon'$ such that if $0 < \epsilon < \epsilon_0$,

$$(8) \quad X_0^\epsilon \subset A'.$$

To show (8), define $A_n := X_0^{\frac{\epsilon'}{n}}$. Since each $A_{n+1} \subset A_n$ are closed by Lemma 13 and $X_0 \subset A_n \subset X_0^{\epsilon'}$ with $X_0^{\epsilon'}$ compact, $\{A_n\}_{n \in \mathbb{N}}$ is a collection of nonempty, nested, compact subsets. Moreover, $\bigcap_{n \in \mathbb{N}} A_n = X_0 \subset A'$ by (7). Thus, by Lemma 17, there exists a positive integer N such that $A_n \subset A'$ if $n \geq N$. In other words, for $\epsilon_0 := \frac{\epsilon'}{N}$, $X_0^\epsilon \subset A'$ if $0 < \epsilon < \epsilon_0$ and each X_0^ϵ is compact since $A' \subset X_0^{\epsilon'}$ is compact as shown earlier. Thus, Claim 20 is proved. \square

REMARK 21. *The assumption that \bar{V} is compact in Theorem 19 can be omitted for metric spaces with the Heine-Borel property, such as when $X = \mathbb{R}^n$.*

3.2. Long-term stability theorem. We are now in the position to show the long-term stability result. In essence, under appropriate conditions, the global error is bounded for all time for conservative methods.

THEOREM 22 (Main stability theorem). *Let $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ with the Euclidean norm $\|\cdot\|$ as their metric. Let $\mathbf{x} \in C^1(I; U) \cap C^p([0, t_{\mu-1}]; U)$ be the unique solution to the ODE (1) with $\mathbf{x}(0) = \mathbf{x}_0$ and $\psi(\mathbf{x}_0) = \mathbf{c}$. For some $\tau_0 > 0$, assume the family of functions $\{\psi^\tau\}_{0 < \tau < \tau_0}$ is equicontinuous on $U \times \cdots \times U \subset \mathbb{R}^{n\mu}$ and let \mathbf{x}_{k+1} be the unique solution to a conservative μ -step discretization (3) of order p with $\psi^\tau(\mathbf{x}_{\mu-1}, \dots, \mathbf{x}_0) = \mathbf{c}^\tau$ and the μ initial values satisfying the hypothesis of Lemma 8.*

If X_0 has a bounded LFN with $\mathbf{x}_0 \in X_0$, then there exists positive constants τ^ and C independent of τ and k such that for all $0 < \tau < \tau^*$ and $k \in \mathbb{N}$,*

$$\|\mathbf{x}(t_k) - \mathbf{x}_k\| \leq C.$$

Proof. Since X_0 is bounded and closed in X by Lemma 13, then X_0 is compact by Heine-Borel's theorem and there is a global solution $\mathbf{x} \in C^1(\mathbb{R}; X_0)$ by Theorem 1. Now the proof proceeds in two main steps.

First, we show for some $\epsilon_0 > 0$, $d_X(X_0^\epsilon, \bigcup_{0 \neq j \in J^\epsilon} X_j^\epsilon) > 0$ for $0 < \epsilon < \epsilon_0$. Recall by Theorem 19, there exists $\epsilon_0 > 0$ so that for $0 < \epsilon < \epsilon_0$, X_0^ϵ is compact and is separated from $\bigcup_{0 \neq j \in J^\epsilon} X_j^\epsilon$. So by Lemma 18, there exists a constant $D_\epsilon > 0$ so that,

$$(9) \quad D_\epsilon \leq \|\mathbf{x} - \mathbf{y}\|, \text{ for } \mathbf{x} \in X_0^\epsilon, \mathbf{y} \in \bigcup_{0 \neq j \in J^\epsilon} X_j^\epsilon.$$

Second, for any fixed $0 < \epsilon < \epsilon_1 \leq \epsilon_0$ where ϵ_1 is a constant to be determined, we prove by induction on $k \in \mathbb{N}$ that $\mathbf{x}_{k+1} \in X_0^\epsilon$ for sufficiently small τ . For $0 \leq k \leq \mu - 1$, by the hypothesis on the initial μ values, we may assume $\{\mathbf{x}_{\mu-1}, \dots, \mathbf{x}_0\} \subset X_0^\epsilon$. Thus, by the strong induction hypothesis on k so that $\{\mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1}\} \subset X_0^\epsilon$, we need to show that there exists a positive constant τ^* such that if $0 < \tau < \tau^*$, $\mathbf{x}_{k+1} \in X_0^\epsilon$ for $0 < \epsilon < \epsilon_1$. For this, we first claim the following:

CLAIM 23. *There exists a constant $\tau_{c,1} > 0$ so that $\mathbf{x}_{k+1} \in U$ for all $0 < \tau < \tau_{c,1}$.*

Define the nested nonempty compact subsets $A_k := X_0^{\frac{\epsilon_0}{k}}$ and so $\bigcap_{k \in \mathbb{N}} A_k = X_0 \subset U$. Thus, by Lemma 17, for some positive integer N , $X_0^\epsilon \subset U$ for all $0 < \epsilon < \epsilon_1 := \frac{\epsilon_0}{N}$. Since $X_0^\epsilon \subset U$ is compact and X_0^ϵ is separated from $X - U$, $L_\epsilon := d_X(X_0^\epsilon, X - U) > 0$ by Lemma 18. As $\{\mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1}\} \subset X_0^\epsilon$, by the local contraction property, there exists $\tau_{c,1}$ depending on $L_\epsilon/2$ and X_0^ϵ so that $\mathbf{x}_{k+1} \in \bigcup_{0 \leq i \leq \mu-1} B_{L_\epsilon/2}(\mathbf{x}_{k-i}) \subset B_{L_\epsilon/2}(X_0^\epsilon) \subset U$ for $0 < \tau < \tau_{c,1}$ and Claim 23 is proved.

Now by Claim 23, $\mathbf{x}_{k+1} \in U$ for all $0 < \tau < \tau_{c,1}$. Combining with equicontinuity and Lemma 5, there exists a $\tau_1 > 0$ such that if $0 < \tau < \tau_1 \leq \min\{\tau_0, \tau_{c,1}\}$,

$$(10) \quad \|\psi(\mathbf{x}_{k+1}) - \psi^\tau(\mathbf{x}_{k+1}, \mathbf{x}_{k+1}, \dots, \mathbf{x}_{k+1})\| < \frac{\epsilon}{3}.$$

Also by Lemma 8, there exists a $\tau_2 > 0$ such that if $0 < \tau < \tau_2 \leq \min\{\tau_0, \tau_{c,1}\}$,

$$(11) \quad \|\mathbf{c} - \mathbf{c}^\tau\| < \frac{\epsilon}{3}.$$

Moreover, by equicontinuity for $0 < \tau < \min\{\tau_0, \tau_{c,1}\}$, there exists a $\delta > 0$ depending only on ϵ such that if $\max_{0 \leq i \leq \mu-1} \|\mathbf{x}_{k+1} - \mathbf{x}_{k-i}\| < \delta$, then

$$(12) \quad \|\psi^\tau(\mathbf{x}_{k+1}, \mathbf{x}_{k+1}, \dots, \mathbf{x}_{k+1}) - \psi^\tau(\mathbf{x}_{k+1}, \mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1})\| < \frac{\epsilon}{3}.$$

Again by the local contraction property, for $\delta_1 := \min\{\delta, \frac{D_\epsilon}{2}\}$, there exists $\tau_{c,2}$ depending on δ and X_0^ϵ so that $\mathbf{x}_{k+1} \in \bigcup_{0 \leq i \leq \mu-1} B_{\delta_1}(\mathbf{x}_{k-i})$ for $0 < \tau < \tau_{c,2}$. I.e. $\max_{0 \leq i \leq \mu-1} \|\mathbf{x}_{k+1} - \mathbf{x}_{k-i}\| < \delta_1$ if $0 < \tau < \tau_{c,2}$. Thus combining (4) with (10)-(12), for $0 < \tau < \tau^* := \min\{\tau_1, \tau_2, \tau_{c,2}\}$,

$$\begin{aligned} \|\psi(\mathbf{x}_{k+1}) - \mathbf{c}\| &\leq \|\psi(\mathbf{x}_{k+1}) - \psi^\tau(\mathbf{x}_{k+1}, \mathbf{x}_{k+1}, \dots, \mathbf{x}_{k+1})\| \\ &\quad + \|\psi^\tau(\mathbf{x}_{k+1}, \mathbf{x}_{k+1}, \dots, \mathbf{x}_{k+1}) - \psi^\tau(\mathbf{x}_{k+1}, \mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1})\| \\ &\quad + \underbrace{\|\psi^\tau(\mathbf{x}_{k+1}, \mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1}) - \mathbf{c}\|}_{=\|\mathbf{c}^\tau - \mathbf{c}\|} < \epsilon. \end{aligned}$$

In other words, we have shown that if $0 < \tau < \tau^*$, $\mathbf{x}_{k+1} \in B_{D_\epsilon/2}(\mathbf{x}_k) \cap \psi^{-1}(\overline{B_\epsilon(\mathbf{c})})$. We now claim that in fact $\mathbf{x}_{k+1} \in X_0^\epsilon$ for $0 < \tau < \tau^*$. Suppose not, then $\mathbf{x}_{k+1} \in$

$\psi^{-1}(\overline{B_\epsilon(\mathbf{c})}) - X_0^\epsilon$. In particular, $\mathbf{x}_{k+1} \in X_j^\epsilon$ for some $j \neq 0$, which leads to a contradiction since by (9),

$$0 < D_\epsilon \leq d(\mathbf{x}_{k+1}, X_0^\epsilon) \leq \underbrace{d(\mathbf{x}_k, X_0^\epsilon)}_{=0} + \underbrace{d(\mathbf{x}_k, \mathbf{x}_{k+1})}_{\leq D_\epsilon/2}.$$

Hence, the induction step is finally proved for any fixed $0 < \epsilon < \epsilon_1$. Since for all $k \in \mathbb{N}$, $\mathbf{x}(t_k) \in X_0 \subset X_0^\epsilon$ and $\mathbf{x}_k \in X_0^\epsilon$ are bounded, we can conclude for $0 < \epsilon < \epsilon_1$,

$$\|\mathbf{x}(t_k) - \mathbf{x}_k\| \leq \sup_{\mathbf{x}, \mathbf{y} \in X_0^\epsilon} \|\mathbf{x} - \mathbf{y}\| = \text{diam}(X_0^\epsilon) =: C. \quad \square$$

REMARK 24. Note that if ψ^τ does not depend on τ explicitly, then the main stability result can be shown readily by using the averaging identity of Corollary 6. Specifically, we have by the averaging identity that,

$$\begin{aligned} \|\psi(\mathbf{x}_{k+1}) - \mathbf{c}\| &= \|\psi^\tau(\mathbf{x}_{k+1}, \dots, \mathbf{x}_{k+1}) - \mathbf{c}\| \\ &\leq \|\psi^\tau(\mathbf{x}_{k+1}, \dots, \mathbf{x}_{k+1}) - \psi^\tau(\mathbf{x}_{k+1}, \mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1})\| \\ &\quad + \|\psi^\tau(\mathbf{x}_{k+1}, \mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1}) - \mathbf{c}\|. \end{aligned}$$

Thus by continuity of ψ^τ and by the local contraction property, $\mathbf{x}_{k+1} \in \psi^{-1}(\overline{B_\epsilon(\mathbf{c})})$ for sufficiently small τ and the proof proceeds similarly as in the equicontinuous case.

REMARK 25. We note that existence of a bounded connected component can be difficult to establish in general, as we discuss in the conclusion. For applications arising from physics, the energy function is a scalar conserved quantity $\psi(\mathbf{x})$ and often satisfies the coercive property; $|\psi(\mathbf{x})| \rightarrow \infty$ as $\|\mathbf{x}\| \rightarrow 0$. In this case, it is well-known that nonempty preimage of ψ is bounded.

REMARK 26. Similarly, it may be difficult to establish in general whether a given bounded connected component of $\psi^{-1}(\mathbf{c})$ has a bounded locally finite neighborhood. Indeed, in the special case when $\psi^{-1}(\mathbf{c})$ has only finite many connected components, it follows from definition that every bounded connected component has a bounded locally finite neighborhood.

3.3. Long-term stability in practice. We conclude this section with a discussion on the effect of error accumulation for conservative methods.

Recall that the main stability result of Theorem 22 holds provided we can guarantee $\mathbf{x}_{k+1} \in B_{D_\epsilon/2}(\mathbf{x}_k) \cap \psi^{-1}(\overline{B_\epsilon(\mathbf{c})})$ for some $\epsilon > 0$ where D_ϵ is the nonzero separation distance between X_0^ϵ and $\bigcup_{0 \neq j \in J^\epsilon} X_j^\epsilon$. However, even for conservative methods, $\psi^\tau\{\mathbf{x}^\tau\}_{k+1} \neq \psi^\tau\{\mathbf{x}^\tau\}_k$ in practice due to round-off error of inexact arithmetic operations in computing ψ^τ . Moreover, \mathbf{x}_{k+1} is often an inexact solution to some iterative method for an implicit method. These errors, while negligible at each time step, can accumulate over long term to be sufficiently large, leading to $\mathbf{x}_{k+1} \notin \psi^{-1}(\overline{B_\epsilon(\mathbf{c})})$ and resulting in the hypotheses of the main theorem being not satisfied. We stress that the error accumulation discussed here is inherent for any finite precision machine when inexact computations are performed over many iterations. In contrast to non-conservative methods, conservative methods are solely limited by these error accumulations depending on machine precision and tolerances used within the method, which we characterize next.

Let ϵ_{mach} be a fixed machine precision and δ_{tol} be a fixed tolerance used as the stopping criterion within the iterative procedure of a given conservative method. Then

on each successive time step t_k , we denote the error E_k accumulated in computing ψ^τ on a finite precision machine as

$$\|\psi^\tau\{\mathbf{x}^\tau\}_k - \psi^\tau\{\mathbf{x}^\tau\}_{k-1}\| \leq E_k(\epsilon_{mach}, \delta_{tol}).$$

Thus by triangle inequality, the error in ψ^τ after N time steps can be estimated as

$$\|\psi^\tau\{\mathbf{x}^\tau\}_N - \mathbf{c}^\tau\| \leq \sum_{k=\mu-1}^N \|\psi^\tau\{\mathbf{x}^\tau\}_k - \psi^\tau\{\mathbf{x}^\tau\}_{k-1}\| \leq \sum_{k=\mu-1}^N E_k(\epsilon_{mach}, \delta_{tol}),$$

where $\mathbf{c}^\tau := \psi^\tau\{\mathbf{x}^\tau\}_{\mu-1}$ from (4). Moreover, suppose E_k can be bounded uniformly by some constant $C_a(\epsilon_{mach}, \delta_{tol}) > 0$, then

$$\|\psi^\tau\{\mathbf{x}^\tau\}_N - \mathbf{c}^\tau\| \leq C_a N.$$

In other words, the error in ψ^τ grows linearly with N in the worst case. However in practice, there may be cancellations within the expressions of the conservative method which can lead to sharper estimates of these round-off errors, as will be illustrated in the numerical example of Section 5. This leads to the following definition of the class of conservative methods with *error accumulation rate* s for some $0 < s \leq 1$.

DEFINITION 27. Fix a machine precision ϵ_{mach} and a tolerance δ_{tol} . A conservative μ -step method \mathbf{F}^τ is said to have an error accumulation rate s if there exists some constants $0 < s(\epsilon_{mach}, \delta_{tol}) \leq 1$ and $C_a(\epsilon_{mach}, \delta_{tol}) > 0$ such that,

$$\|\psi^\tau\{\mathbf{x}^\tau\}_N - \mathbf{c}^\tau\| \leq C_a N^s.$$

Now we state the **arbitrarily long-term stability** theorem on finite precision machines.

THEOREM 28. Suppose the hypotheses of Theorem 22 (Main stability theorem) are satisfied and the μ -step method \mathbf{F}^τ has an error accumulation rate s . If X_0 has a bounded LFN with $\mathbf{x}_0 \in X_0$, then there exists a positive integer N_{max} depending on ϵ_{mach} and δ_{tol} , a positive constant C independent of N_{max} and a positive constant τ^* independent of τ such that for all $0 < \tau < \tau^*$ and $0 \leq k \leq N_{max}$,

$$\|\mathbf{x}(t_k) - \mathbf{x}_k\| \leq C.$$

Proof. We highlight the differences in the proof, as it is nearly identical to the proof of Theorem 22. For brevity, we shall focus on the case when ψ^τ does not explicitly depend on τ , as the same conclusion follows for the equicontinuous case (with possibly a smaller N_{max}). Since the quantity $\|\psi^\tau\{\mathbf{x}^\tau\}_k - \mathbf{c}^\tau\| \neq 0$ with finite precision arithmetic, by Remark 24, we need to instead establish the following estimate,

$$(13) \quad \|\psi^\tau\{\mathbf{x}^\tau\}_{k+1} - \mathbf{c}\| \leq \|\psi^\tau\{\mathbf{x}^\tau\}_{k+1} - \mathbf{c}^\tau\| + \|\mathbf{c}^\tau - \mathbf{c}\| \leq \frac{\epsilon}{2},$$

for $0 < \epsilon < \epsilon_0$ where ϵ_0 is the largest radius around \mathbf{c} for which X_0 is separated from the other connected components X_j contained in $\psi^{-1}(\overline{B_{\epsilon_0}(\mathbf{c})})$. By Lemma 8, $\|\mathbf{c}^\tau - \mathbf{c}\| \leq \frac{\epsilon}{4}$ for sufficiently small τ . Moreover, since \mathbf{F}^τ is assumed to have an error accumulation rate s , $\|\psi^\tau\{\mathbf{x}^\tau\}_{k+1} - \mathbf{c}^\tau\| \leq C_a N_{max}^s$, for all $k+1 \leq N_{max}$. Thus, the estimate (13) follows if $N_{max} := \left(\frac{\epsilon}{4C_a}\right)^{1/s}$. Finally, we also note that $C := \text{diam}(X_0^\epsilon)$ as in the main theorem and is independent of N_{max} . \square

4. Conservative method for autonomous ODEs. We now shift our discussion from the theory presented in Sections 2 and 3 to applications. In this section, we discuss constructions of conservative methods. Specifically, in this section, we derive sufficient conditions for constructing conservative methods based on the recent multiplier method [25]. Moreover, in the case when conserved quantities are of polynomial form, we give explicit formulas for first order conservative discretization using the multiplier method approach.

4.1. Background on the multiplier method. We briefly review the multiplier method [25] in the special case of autonomous system (1) with $1 \leq m \leq n$ linearly independent² conserved quantities ψ on some open subset $U \subset \mathbb{R}^n$. The essential idea of the multiplier method is that conserved quantities of ODEs/PDEs are associated with so-called *characteristics of conservation laws* or *conservation law multipliers* [21, 3], so that when multiplied with the equations yields a divergence expression representing the conserved quantities. The multiplier method proposes to discretize the multipliers and conserved quantities in a manner such that the divergence theorem is preserved discretely, which implies the conservative property 4 holds. In the present paper, we focus on conserved quantities and conservation law multipliers depending only on \mathbf{x} and assume the following hypothesis:

HYPOTHESIS 29. *Given a system of ODEs (1) with conserved quantities $\psi(\mathbf{x})$, there exists a corresponding multiplier matrix $\Lambda : \mathbb{R}^n \rightarrow L(\mathbb{R}^n, \mathbb{R}^m)$ such that,*

$$(14) \quad \Lambda(\mathbf{x})\mathbf{F}[\mathbf{x}] = \Lambda(\mathbf{x})(\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x})) = D_t\psi(\mathbf{x}), \text{ for all } \mathbf{x} \in C^1(I; U),$$

where D_t is the total derivative with respect to t .

It turns out that the case with $m = n$ only arises for the trivial system when $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ for all $\mathbf{x} \in U$, as the following Lemma 30 will show. This, however, does not preclude the autonomous system (1) from possessing n linear independent conserved quantities, since it is still possible to have conserved quantities and conservation law multipliers which depend on higher order derivatives such as $\psi(\mathbf{x}, \dot{\mathbf{x}})$. Using hypothesis 29, we can deduce a simple necessary condition relating ψ , Λ and \mathbf{f} .

LEMMA 30. *If ψ is continuously differentiable on U , then for all $\mathbf{y} \in U$,*

$$(15) \quad \mathbf{J}_\psi(\mathbf{y}) = \Lambda(\mathbf{y}),$$

$$(16) \quad \Lambda(\mathbf{y})\mathbf{f}(\mathbf{y}) = \mathbf{0},$$

where $\mathbf{J}_\psi(\mathbf{x})$ is the Jacobian matrix of ψ at \mathbf{x} .

Proof. Let $\mathbf{y} \in U$. By (14) and the chain rule, for any $\mathbf{x} \in C^1(I; U)$,

$$(17) \quad \Lambda(\mathbf{x})(\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x})) = D_t\psi(\mathbf{x}) = \mathbf{J}_\psi(\mathbf{x})\dot{\mathbf{x}}$$

Defining the constant function $\mathbf{x}(t) = \mathbf{y}$ in (17) implies (16). Let \mathbf{e}_i be the i -th standard basis vector for each $i = 1, \dots, n$. By (16), evaluating $\mathbf{x}(t) = \mathbf{e}_i(t - t_0) + \mathbf{y}$ at $t = t_0 \in I$ in (17) and combining with (16) implies (15) for the i -th column vector. \square

REMARK 31. *If $m = n$ or equivalently Λ is a square matrix, then $\psi(\mathbf{x})$ being linearly independent on U implies $\mathbf{J}_\psi(\mathbf{x}) = \Lambda(\mathbf{x})$ has full rank on U by (15). So it follows from (16) that $\mathbf{f}(\mathbf{x})$ must be the zero vector for all $\mathbf{x} \in U$. In other words, for a given non-trivial autonomous system, there is at most $n - 1$ linearly independent conserved quantities of the form $\psi(\mathbf{x})$.*

²By linear independence, we mean the Jacobian matrix $\mathbf{J}_\psi(\mathbf{x})$ has maximal rank for all $\mathbf{x} \in U$.

As a consequence, if the discretized multiplier Λ^τ is consistent to Λ , then there must an $m \times m$ minor of Λ^τ which is invertible for sufficiently small τ .

THEOREM 32. *Suppose the discretized multiplier matrix $\Lambda^\tau : \mathbb{R}^{n\mu} \rightarrow L(\mathbb{R}^n, \mathbb{R}^m)$ is consistent of order p to the multiplier Λ ; that is for $\mathbf{x} \in C^p(I; U)$ and each time step t_k , there exists a positive constant C_Λ independent of τ and such that for $I_k = [t_{k-\mu+1}, t_k]$,*

$$\|\Lambda(\mathbf{x}) - \Lambda^\tau(\mathbf{x}(t_k), \dots, \mathbf{x}(t_{k-\mu+1}))\|_{op} \leq C_\Lambda \|\mathbf{x}\|_{C^p(I_k)} \tau^p.$$

For some $\tau_0 > 0$, assume Λ^τ is continuous with respect to τ for all $0 < \tau < \tau_0$ and the family of discrete multipliers $\{\Lambda^\tau\}_{0 < \tau < \tau_0}$ is equicontinuous on $U \times \dots \times U \subset \mathbb{R}^{n\mu}$. Also, denote $\{\mathbf{x}_k : k \in \mathbb{N}\}$ as the sequence generated by solving the solution of a μ -step discretization (3) with the local contraction property. Then for some $\tau^* > 0$, if $0 < \tau < \tau^*$, there is an $m \times m$ minor of Λ^τ , denoted as $\tilde{\Lambda}^\tau(\mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1})$, that is invertible and the operator norm of its inverse is bounded above independent of τ .

Proof. The proof is broken into three main steps. First, since ψ is linearly independent on U , by the relation (15), there must be an $m \times m$ minor of Λ denoted as $\tilde{\Lambda}$ with $\det(\tilde{\Lambda}) \neq 0$ on U . Denote the corresponding $m \times m$ minor of Λ^τ as $\tilde{\Lambda}^\tau$.

Second, since $\{\Lambda^\tau\}_{0 < \tau < \tau_0}$ is equicontinuous, by a very similar proof as in Theorem (5), one can show that for all $\mathbf{y} \in U$, $\tilde{\Lambda}(\mathbf{y}) = \lim_{\tau \rightarrow 0} \tilde{\Lambda}^\tau(\mathbf{y}, \dots, \mathbf{y})$. Thus combining with the fact that the determinant function is continuous, we have that for all $\mathbf{y} \in U$,

$$0 \neq \det(\tilde{\Lambda}(\mathbf{y})) = \lim_{\tau \rightarrow 0} \det(\tilde{\Lambda}^\tau(\mathbf{y}, \dots, \mathbf{y})).$$

In other words, for some $\tau_1 > 0$, if $0 < \tau < \min\{\tau_0, \tau_1\}$ and $\mathbf{y} \in U$,

$$(18) \quad \det(\tilde{\Lambda}^\tau(\mathbf{y}, \dots, \mathbf{y})) \neq 0.$$

Finally, by equicontinuity again and local contraction property of the μ -step method, for any $\epsilon > 0$ there is a $\tau_2 > 0$ such that if $0 < \tau < \tau_2$,

$$(19) \quad \left\| \tilde{\Lambda}^\tau(\mathbf{x}_k, \dots, \mathbf{x}_k) - \tilde{\Lambda}^\tau(\mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1}) \right\|_{op} < \epsilon.$$

Combining (19) with (18) and again by continuity of the determinant, it follows that for some positive constant γ , $|\det(\tilde{\Lambda}^\tau(\mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1}))| \geq \gamma > 0$, for all $0 < \tau < \tau^* := \min\{\tau_0, \tau_1, \tau_2\}$.

To show that the operator norm of its inverse is bounded above independent of τ , it follows from the adjugate formula of the inverse matrix that for $0 < \tau < \tau^*$,

$$\begin{aligned} \left\| (\tilde{\Lambda}^\tau)^{-1}(\mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1}) \right\|_{op} &= \frac{\left\| \text{adj}(\tilde{\Lambda}^\tau)(\mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1}) \right\|_{op}}{|\det(\tilde{\Lambda}^\tau(\mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1}))|} \\ &\leq \frac{1}{\gamma} \sup_{\tau \in (0, \tau^*)} \left\| \text{adj}(\tilde{\Lambda}^\tau)(\mathbf{x}_k, \dots, \mathbf{x}_{k-\mu+1}) \right\|_{op}. \quad \square \end{aligned}$$

REMARK 33. *Analogous to Remark 24 at the end of Theorem 22, the equicontinuity condition of Λ^τ can be replaced with the condition that Λ^τ do not depend on τ explicitly, in which case an analogous proof using an averaging identity for Λ^τ follows in a similar fashion; we omit the details of the proof here.*

4.2. Sufficient conditions for conservative discretizations. Now we apply the consistency theorem of the multiplier method to the autonomous system (1).

THEOREM 34 (Theorem 6 of [25]). *Let $\psi(\mathbf{x})$ be $1 \leq m \leq n-1$ linearly independent conserved quantities on U for the autonomous system (1) with the associated multiplier matrix $\Lambda(\mathbf{x})$. Without loss of generality (with reordering of equations if necessary), denote $\Lambda = \begin{pmatrix} \tilde{\Lambda} & \Sigma \end{pmatrix}$ where $\tilde{\Lambda}(\mathbf{x})$ is the $m \times m$ minor which is invertible on U . Also, partition the associated equations as $\mathbf{F} = \begin{pmatrix} \mathbf{F} & \mathbf{G} \end{pmatrix}^T$. Suppose Λ^τ is a consistent discrete multiplier matrix of order p , denoted as $\Lambda^\tau = \begin{pmatrix} \tilde{\Lambda}^\tau & \Sigma^\tau \end{pmatrix}$. Also, suppose the discrete functions $D_t^\tau \psi : \mathbb{R}^{n\mu} \rightarrow \mathbb{R}^m$ and $\mathbf{G}^\tau : \mathbb{R}^{n\mu} \rightarrow \mathbb{R}^{n-m}$ are consistent to $D_t \psi$ and \mathbf{G} of order p . Then for sufficiently small τ , the inverse of $\tilde{\Lambda}^\tau$ exists and the following discretization \mathbf{F}^τ is consistent to \mathbf{F} of order p :*

$$(20) \quad \mathbf{F}^\tau \{\mathbf{x}^\tau\} := \begin{pmatrix} \left(\tilde{\Lambda}^\tau \{\mathbf{x}^\tau\} \right)^{-1} (D_t^\tau \psi \{\mathbf{x}^\tau\} - \Sigma^\tau \{\mathbf{x}^\tau\} \mathbf{G}^\tau \{\mathbf{x}^\tau\}) \\ \mathbf{G}^\tau \{\mathbf{x}^\tau\} \end{pmatrix}.$$

Proof. This directly follows from the consistency result of Theorem 6 of [25] applied to the present case of autonomous system. We note that Theorem 32 implies the inverse of $\tilde{\Lambda}^\tau$ exists for sufficiently small τ . \square

Analogous to Lemma 30, we now state a sufficient condition for constructing conservative discretizations for (1).

THEOREM 35. *Assume the hypothesis of Theorem 34 holds and suppose the discrete functions $\psi^\tau, \Lambda^\tau, D_t^\tau \mathbf{x}, \mathbf{f}^\tau$ are consistent of order p to $\psi, \Lambda, \dot{\mathbf{x}}, \mathbf{f}$, respectively. Further assume that at each $k = 1, 2, \dots$,*

$$(21) \quad \frac{\psi^\tau \{\mathbf{x}^\tau\}_k - \psi^\tau \{\mathbf{x}^\tau\}_{k-1}}{\tau} = \Lambda^\tau \{\mathbf{x}^\tau\}_k D_t^\tau \mathbf{x}_k,$$

$$(22) \quad \Lambda^\tau \{\mathbf{x}^\tau\}_k \mathbf{f}^\tau \{\mathbf{x}^\tau\}_k = \mathbf{0}.$$

Then the following discretization is consistent to \mathbf{F} of order p ,

$$(23) \quad \mathbf{F}^\tau \{\mathbf{x}^\tau\} := D_t^\tau \mathbf{x} - \mathbf{f} \{\mathbf{x}^\tau\} = \mathbf{0},$$

and the solution of (23) satisfies $\psi^\tau \{\mathbf{x}^\tau\}_k = \psi^\tau \{\mathbf{x}^\tau\}_{k-1}$ for all $k = 1, 2, \dots$.

Proof. Write $\mathbf{f}^\tau = \begin{pmatrix} \tilde{\mathbf{f}}^\tau & \hat{\mathbf{f}}^\tau \end{pmatrix}^T$ where $\tilde{\mathbf{f}}^\tau, \hat{\mathbf{f}}^\tau$ is the first m components and last $n-m$ components of \mathbf{f}^τ , respectively. Similarly for $D_t^\tau \mathbf{x} = \begin{pmatrix} D_t^\tau \tilde{\mathbf{x}} & D_t^\tau \hat{\mathbf{x}} \end{pmatrix}^T$. By (20) at each $k = 1, 2, \dots$, with $\mathbf{G}^\tau = D_t^\tau \hat{\mathbf{x}} - \hat{\mathbf{f}}^\tau$ and $D_t^\tau \psi \{\mathbf{x}^\tau\} = \frac{1}{\tau} (\psi^\tau \{\mathbf{x}^\tau\}_k - \psi^\tau \{\mathbf{x}^\tau\}_{k-1})$,

$$\begin{aligned} \tilde{\mathbf{F}}^\tau \{\mathbf{x}^\tau\}_k &= (\tilde{\Lambda}^\tau \{\mathbf{x}^\tau\}_k)^{-1} \left(\frac{\psi^\tau \{\mathbf{x}^\tau\}_k - \psi^\tau \{\mathbf{x}^\tau\}_{k-1}}{\tau} - \Sigma^\tau \{\mathbf{x}^\tau\}_k (D_t^\tau \hat{\mathbf{x}} - \hat{\mathbf{f}}^\tau \{\mathbf{x}^\tau\}_k) \right) \\ &= (\tilde{\Lambda}^\tau \{\mathbf{x}^\tau\}_k)^{-1} \left(\Lambda^\tau \{\mathbf{x}^\tau\}_k D_t^\tau \mathbf{x}_k - \Sigma^\tau \{\mathbf{x}^\tau\}_k D_t^\tau \hat{\mathbf{x}} + \Sigma^\tau \{\mathbf{x}^\tau\}_k \hat{\mathbf{f}}^\tau \{\mathbf{x}^\tau\}_k \right) \quad \text{by (21)} \\ &= (\tilde{\Lambda}^\tau \{\mathbf{x}^\tau\}_k)^{-1} \left(\tilde{\Lambda}^\tau \{\mathbf{x}^\tau\}_k D_t^\tau \tilde{\mathbf{x}}_k + \Sigma^\tau \{\mathbf{x}^\tau\}_k D_t^\tau \hat{\mathbf{x}}_k \right. \\ &\quad \left. - \Sigma^\tau \{\mathbf{x}^\tau\}_k D_t^\tau \hat{\mathbf{x}}_k - \tilde{\Lambda}^\tau \{\mathbf{x}^\tau\}_k \tilde{\mathbf{f}}^\tau \{\mathbf{x}^\tau\}_k \right) \quad \text{by (22)} \\ &= D_t^\tau \tilde{\mathbf{x}}_k - \tilde{\mathbf{f}}^\tau \{\mathbf{x}^\tau\}_k \end{aligned}$$

Thus, by Theorem 7 and Corollary 3 of [25], \mathbf{F}^τ is of order p and the solution to (23) satisfies $\psi^\tau \{\mathbf{x}^\tau\}_k = \psi^\tau \{\mathbf{x}^\tau\}_{k-1}$ for all $k = 1, 2, \dots$. \square

REMARK 36. In the case of first order approximations, we note that conditions (21) and (22) is equivalent to the discrete gradient approximation and Lemma 2.2 of [11].

4.3. Conserved quantities of polynomial form. For conserved quantities of polynomial form, it turns out one can always construct first order Λ^τ , $D_t^\tau \mathbf{x}_k$ and \mathbf{f}^τ satisfying conditions (21) and (22). First, we need the following lemma.

LEMMA 37. Suppose $g^\tau(\mathbf{x}_k)$ is a degree N polynomial of $\mathbf{x}_k = (x_{1,k}, \dots, x_{n,k})^T$,

$$g^\tau(\mathbf{x}_k) = \sum_{|\alpha| \leq N} c_\alpha \mathbf{x}_k^\alpha = \sum_{|\alpha| \leq N} c_\alpha \prod_{r=1}^n x_{r,k}^{\alpha_r},$$

for some coefficient c_α indexed by the multi-index α . Then the forward difference of $g^\tau(\mathbf{x}_k)$ can be expressed as,

$$\begin{aligned} & \frac{g^\tau(\mathbf{x}_k) - g^\tau(\mathbf{x}_{k-1})}{\tau} \\ &= \sum_{j=1}^n \left(\frac{x_{j,k} - x_{j,k-1}}{\tau} \right) \left(\sum_{|\alpha| \leq N} c_\alpha \left(\sum_{l=0}^{\alpha_j-1} x_{j,k}^l x_{j,k-1}^{\alpha_j-l-1} \right) \prod_{r=1}^{j-1} x_{r,k-1}^{\alpha_r} \prod_{s=j+1}^n x_{s,k}^{\alpha_s} \right). \end{aligned}$$

Proof. By induction on n , it can be shown that,

$$\begin{aligned} \prod_{r=1}^n x_{r,k}^{\alpha_r} - \prod_{r=1}^n x_{r,k-1}^{\alpha_r} &= \sum_{j=1}^n \left(x_{j,k}^{\alpha_j} - x_{j,k-1}^{\alpha_j} \right) \prod_{r=1}^{j-1} x_{r,k-1}^{\alpha_r} \prod_{s=j+1}^n x_{s,k}^{\alpha_s} \\ &= \sum_{j=1}^n (x_{j,k} - x_{j,k-1}) \left(\sum_{l=0}^{\alpha_j-1} x_{j,k}^l x_{j,k-1}^{\alpha_j-l-1} \right) \prod_{r=1}^{j-1} x_{r,k-1}^{\alpha_r} \prod_{s=j+1}^n x_{s,k}^{\alpha_s}. \end{aligned}$$

where the last equality follows from geometric sum. The result now follows since,

$$\begin{aligned} \frac{g^\tau(\mathbf{x}_k) - g^\tau(\mathbf{x}_{k-1})}{\tau} &= \sum_{|\alpha| \leq N} \frac{c_\alpha}{\tau} \left(\prod_{r=1}^n x_{r,k}^{\alpha_r} - \prod_{r=1}^n x_{r,k-1}^{\alpha_r} \right) \\ &= \sum_{|\alpha| \leq N} \frac{c_\alpha}{\tau} \left(\sum_{j=1}^n (x_{j,k} - x_{j,k-1}) \left(\sum_{l=0}^{\alpha_j-1} x_{j,k}^l x_{j,k-1}^{\alpha_j-l-1} \right) \prod_{r=1}^{j-1} x_{r,k-1}^{\alpha_r} \prod_{s=j+1}^n x_{s,k}^{\alpha_s} \right) \\ &= \sum_{j=1}^n \left(\frac{x_{j,k} - x_{j,k-1}}{\tau} \right) \left(\sum_{|\alpha| \leq N} c_\alpha \left(\sum_{l=0}^{\alpha_j-1} x_{j,k}^l x_{j,k-1}^{\alpha_j-l-1} \right) \prod_{r=1}^{j-1} x_{r,k-1}^{\alpha_r} \prod_{s=j+1}^n x_{s,k}^{\alpha_s} \right). \square \end{aligned}$$

Combining Theorem 35 and Lemma 37 gives the following result.

THEOREM 38. Let $\mathbf{F}[\mathbf{x}] := \dot{\mathbf{x}} - \mathbf{f}(\mathbf{x})$ be a first order autonomous system with $1 \leq m \leq n-1$ linearly independent conserved quantities $\psi(\mathbf{x})$ of degree N polynomials, i.e. ψ has components of the form,

$$\psi_i(\mathbf{x}) = \sum_{|\alpha| \leq N} c_{i,\alpha} \mathbf{x}^\alpha, \quad i = 1, \dots, m.$$

Define the discrete multiplier matrix $\Lambda^\tau(\mathbf{x}_k)$ for $1 \leq i \leq m$ and $1 \leq j \leq n$ as

$$(24) \quad (\Lambda^\tau(\mathbf{x}_k))_{ij} = \sum_{|\alpha| \leq N} c_{i,\alpha} \left(\sum_{l=0}^{\alpha_j-1} x_{j,k}^l x_{j,k-1}^{\alpha_j-l-1} \right) \prod_{r=1}^{j-1} x_{r,k-1}^{\alpha_r} \prod_{s=j+1}^n x_{s,k}^{\alpha_s}.$$

Denote $\Lambda^\tau = (\tilde{\Lambda}^\tau \quad \Sigma^\tau)$ with $\tilde{\Lambda}^\tau$ of size $m \times m$ and define the discretization:

$$(25) \quad \mathbf{F}^\tau \{\mathbf{x}^\tau\}_k := \frac{\mathbf{x}_k - \mathbf{x}_{k-1}}{\tau} - \begin{pmatrix} -(\tilde{\Lambda}^\tau(\mathbf{x}_k))^{-1} \Sigma^\tau(\mathbf{x}_k) \\ I_{n-m} \end{pmatrix} \hat{\mathbf{f}}^\tau \{\mathbf{x}^\tau\}_k = \mathbf{0}.$$

where $\hat{\mathbf{f}}^\tau$ is at least first order discretization of $\hat{\mathbf{f}}$. Then the following holds.

1. For some $\tau_0 > 0$, the inverse of $\tilde{\Lambda}^\tau(\mathbf{x}_k)$ exists for $0 < \tau \leq \tau_0$.
2. \mathbf{F}^τ is a first order accurate discretization to \mathbf{F} .
3. The solution to (25) satisfies $\psi^\tau(\mathbf{x}_k) = \psi^\tau(\mathbf{x}_{k-1})$ where for $1 \leq i \leq n$,

$$\psi_i^\tau(\mathbf{x}_k) = \sum_{|\alpha| \leq N} c_{i,\alpha} x_k^\alpha.$$

Proof. 1 follows from Theorem 32. To show 2 and 3, we verify the two conditions of Theorem 35. Define $D_t^\tau \mathbf{x}_k := \frac{\mathbf{x}_k - \mathbf{x}_{k-1}}{\tau}$ and Lemma 37 with $g^\tau(\mathbf{x}_k) = \psi_i^\tau(\mathbf{x}_k)$ for each $i = 1, \dots, m$ implies,

$$\begin{aligned} & (\Lambda^\tau \{\mathbf{x}^\tau\}_k D_t^\tau \mathbf{x}_k)_i \\ &= \sum_{j=1}^n \left(\sum_{|\alpha| \leq N} c_{i,\alpha} \left(\sum_{l=0}^{\alpha_j-1} x_{j,k}^l x_{j,k-1}^{\alpha_j-l-1} \right) \prod_{r=1}^{j-1} x_{r,k-1}^{\alpha_r} \prod_{s=j+1}^n x_{s,k}^{\alpha_s} \right) \left(\frac{x_{j,k} - x_{j,k-1}}{\tau} \right) \\ &= \frac{\psi_i^\tau \{\mathbf{x}^\tau\}_k - \psi_i^\tau \{\mathbf{x}^\tau\}_{k-1}}{\tau}, \end{aligned}$$

which verifies the condition of (21). Moreover, (22) is also satisfied, since

$$\Lambda^\tau \{\mathbf{x}^\tau\}_k \mathbf{f}^\tau \{\mathbf{x}^\tau\}_k = \underbrace{\left(\tilde{\Lambda}^\tau \{\mathbf{x}^\tau\}_k \quad \Sigma^\tau \{\mathbf{x}^\tau\}_k \right) \begin{pmatrix} -(\tilde{\Lambda}^\tau \{\mathbf{x}^\tau\}_k)^{-1} \Sigma^\tau \{\mathbf{x}^\tau\}_k \\ I_{n-m} \end{pmatrix} \hat{\mathbf{f}}^\tau \{\mathbf{x}^\tau\}_k}_{=:\mathbf{f}^\tau \{\mathbf{x}^\tau\}_k} = \mathbf{0}.$$

By Theorem 35, the proposed discretization \mathbf{F}^τ is consistent with \mathbf{F} to first order, as $D_t^\tau \mathbf{x}_k$ and Λ^τ are consistent with $\dot{\mathbf{x}}, \Lambda$ to first order. \square

5. Numerical example: Elliptic curve. We now illustrate the long-term stability theorem with an example and the use of (25). For any $a \in \mathbb{R}$, consider

$$(26) \quad \mathbf{F}[\mathbf{x}] := \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} - \begin{pmatrix} 2y \\ 3x^2 + a \end{pmatrix} = \mathbf{0}, \quad \mathbf{x}(0) = \mathbf{x}_0.$$

Multiplying \mathbf{F} by the multiplier matrix $\Lambda(\mathbf{x}) = \begin{pmatrix} -3x^2 - a & 2y \end{pmatrix}$, shows that

$$(27) \quad \psi(\mathbf{x}) := y^2 - x^3 - ax$$

is a conserved quantity. Indeed, if \mathbf{x} is the unique solution to (26),

$$0 = \Lambda(\mathbf{x}) \mathbf{F}[\mathbf{x}] = (-3x^2 - a)(\dot{x} - 2y) + 2y(\dot{y} - 3x^2 - a) = D_t \psi(\mathbf{x})$$

In particular, (27) implies $\psi(\mathbf{x}) = b$ is an elliptic curve for any $b \in \mathbb{R}$. Since $\psi(\mathbf{x})$ is a cubic polynomial, (24) implies the discrete multiplier matrix,

$$\Lambda^\tau \{\mathbf{x}^\tau\}_k = \begin{pmatrix} -x_{k-1}^2 - x_{k-1}x_k - x_k^2 - a & y_{k-1} + y_k \end{pmatrix}.$$

Let $\tilde{\Lambda}^\tau \{\mathbf{x}^\tau\}_k = (-x_{k-1}^2 - x_{k-1}x_k - x_k^2 - a)$, $\Sigma^\tau \{\mathbf{x}^\tau\}_k = (y_{k-1} + y_k)$ and $\hat{\mathbf{f}}^\tau \{\mathbf{x}^\tau\}_k = (x_{k-1}^2 + x_{k-1}x_k + x_k^2 + a)$. Then Theorem 38 gives the first order conservative discretization for (26) which preserves $\psi^\tau \{\mathbf{x}^\tau\}_k = y_k^2 - x_k^3 - ax_k$:

$$(28) \quad \mathbf{F}^\tau \{\mathbf{x}^\tau\}_k := \begin{pmatrix} \frac{x_k - x_{k-1}}{\tau} \\ \frac{y_k - y_{k-1}}{\tau} \end{pmatrix} - \begin{pmatrix} y_{k-1} + y_k \\ x_{k-1}^2 + x_{k-1}x_k + x_k^2 + a \end{pmatrix} = \mathbf{0}$$

REMARK 39. We note that (28) can also be derived using the average vector field method [23], since closed form integration can be performed for polynomials.

For any $a, b \in \mathbb{R}^n$, it is well-known that the elliptic curve $\psi(\mathbf{x}) = b$ has at most two connected components. In particular, if the sign of the discriminant of the cubic polynomial $p(\mathbf{x}) := \psi(\mathbf{x}) - b$ is given by $\Delta(p) := 4a^3 + 27b^2$ is negative, the elliptic curve has two connected components with one bounded and the other unbounded. Otherwise, the elliptic curve has only an unbounded connected component if $\Delta(p) > 0$.

We compare numerical results of the conservative first order method of (28) with standard first order explicit/implicit methods (Euler/Backward Euler) and symplectic second order explicit/implicit methods (Störmer-Verlet/Midpoint) [14]. We considered the case of two connected components with $a = -1$ and $b = 0.3849$ ($\Delta(p) < 0$), where the bounded connected component of X_0 and unbounded connected component of X_1 are close to each other but separated as shown in Figure 1a and 1b. In all five methods, the initial conditions were set to be $x_0 = 0.571$ and $y_0 = \sqrt{x_0^3 + ax_0 + b} \approx 8.33 \times 10^{-3}$, which implies, for τ sufficiently small, the exact solution should remain within the bounded connected component of X_0 . We have used an uniform time step size of $\tau = 0.3$ with $N = 5 \times 10^3$ time steps and we employed an absolute tolerance of $\delta_{tol} = 5 \times 10^{-16}$ with a maximum of 50 Newton's iterations per time step for the implicit methods.

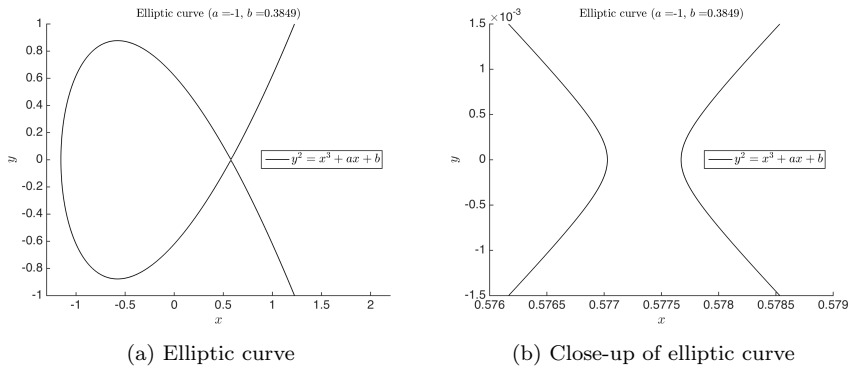


FIG. 1. Two connected components of the preimage of $\psi^{-1}(\{b\})$.

Figure 2 shows that Euler method gives an unbounded solution and Backward Euler method leads to a decaying solution to a fixed point $\mathbf{x}^* = (-1/\sqrt{3}, 0)^T$. Figure 2

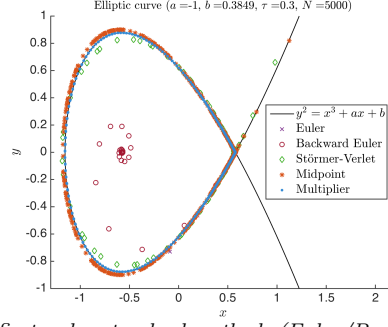


FIG. 2. Comparison of first order standard methods (Euler/Backward Euler), second order symplectic methods (Störmer-Verlet/Midpoint) and first order conservative method (Multiplier method).

and 3a shows Störmer-Verlet method gives a solution which loops around the bounded connected component of X_0 once before exiting to the unbounded connected component of X_1 . Similarly, Figure 2 and 3b shows the solution of the Midpoint method loops around X_0 longer than the Störmer-Verlet method before eventually exiting to X_1 . In contrast, all three figures show that the first order multiplier method gives a solution which remains essentially on the bounded connect component of X_0 and indeed we observed an error in ψ of $\max_{1 \leq i \leq 5 \times 10^3} |\psi(\mathbf{x}_i) - b| \sim 6.6 \times 10^{-15}$.

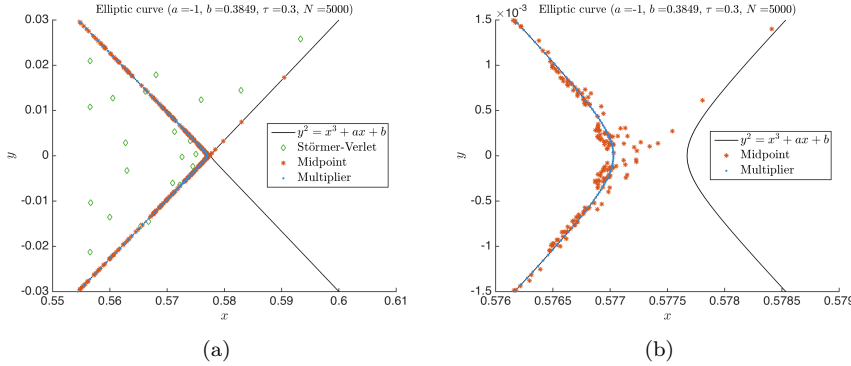
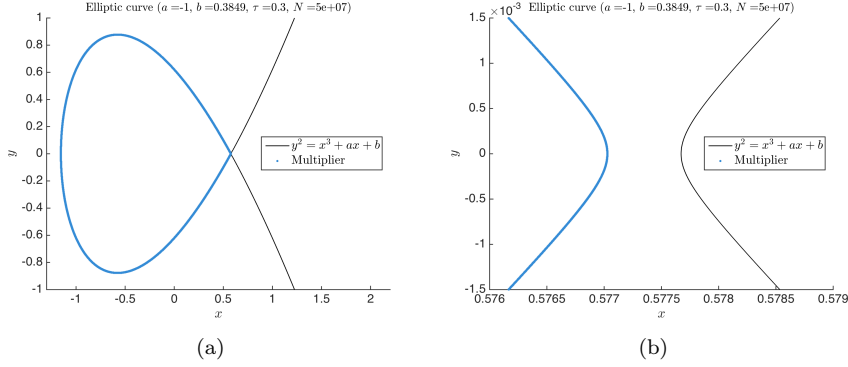
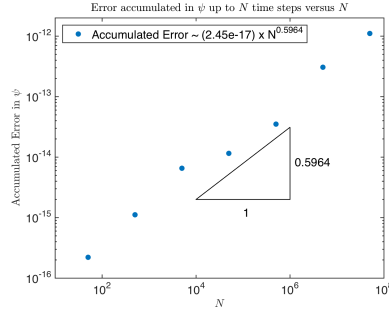


FIG. 3. Close-up comparison between the Störmer-Verlet, Midpoint and Multiplier method in Figure 3a and the Midpoint and Multiplier method in Figure 3b.

Next, we increase the number of time steps to $N = 5 \times 10^7$ while fixing all other parameters. As shown in Figures 3a and 3b, the first order multiplier method again gives a solution which stays near the bounded connected component of X_0 . As we increase the number of time steps, we expect an increase of the error in ψ due to round-off error accumulation and inexact iterative solutions as discussed in Section 3.3. Indeed, we observed the error in ψ now to be $\max_{1 \leq i \leq 5 \times 10^7} |\psi(\mathbf{x}_i) - b| \approx 1.1 \times 10^{-12}$.

To investigate further on error accumulation of ψ as N increases, Figure 5 shows a log-log plot of the accumulated error $\max_{1 \leq i \leq N} |\psi(\mathbf{x}_i) - b|$ for various N . By a linear regression, the accumulated error $E(N)$ was estimated to be $E \approx (2.45 \times 10^{-17}) \times N^{0.5964}$, where the error accumulation rate of 0.5964 is due to inherent round-off cancellations within the conservative discretization.

FIG. 4. *First order conservative method for $N = 5 \times 10^7$.*FIG. 5. *Error accumulation in ψ versus N .*

To estimate the maximum number of time steps N_{max} as stipulated in Theorem 28, we need the largest $\epsilon > 0$ such that $\psi^{-1}((b - \epsilon, b + \epsilon))$ still has two connected components. In the present case of elliptic curve, we know that the two connected components merges into one precisely when the discriminant $\Delta(p)$ changes sign. Thus, computing $\Delta(p) = 0$ gives $\epsilon \approx 1.8 \times 10^{-7}$, which implies a maximum number of time steps $N_{max} = \left(\frac{\epsilon}{4C_a}\right)^{1/s} \approx 6.9 \times 10^{16}$ so that the conservative method can be guaranteed to have a bounded global error. This is in stark contrast to previous four methods (some of which are second order) where their solutions decay to a fixed point or become unbounded for $N \leq 5 \times 10^3$.

REMARK 40. *We elected to not make comparison with projection-based conservative methods; methods which first evolve in time using traditional methods and after some time period project the discrete solution back onto the constraint of conserved quantities. While this approach can make any traditional method conservative, we note the long-term stability result may not hold for these methods, as the composition of evolution and projection may not satisfy the local contraction property.*

6. Conclusion. In this paper, we have presented a long-term stability result for conservative methods in the case of autonomous ODEs; specifically the global error is, in principle, bounded for all time. On finite precision machines, the global error is shown to be bounded up to some arbitrarily long time depending only on machine precision and tolerance. Since the main idea is mostly based on topological ideas, we

believe the stability result can be generalized to certain non-autonomous ODEs and PDEs; such is the goal of our current work.

Acknowledgments. ATSW would like to thank Siddarth Sankaran for our discussions on algebraic geometry related to this work and Chris Budd for pointing out the connection of our polynomial formula with the average vector field method.

REFERENCES

- [1] D. N. ARNOLD, R. S. FALK, AND R. WINTHER, *Finite element exterior calculus, homological techniques, and applications*, Acta Numerica, (2006), pp. 1–155.
- [2] G. BENETTIN AND A. GIORGILLI, *On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms*, J. Statist. Phys., 74 (1994), pp. 1117–1143.
- [3] G. BLUMAN, A. CHEVIAKOV, AND S. ANCO, *Applications of Symmetry Methods to Partial Differential Equations*, Springer, 2010.
- [4] P. B. BOCHEV AND J. M. HYMAN, *Principles of mimetic discretizations of differential operators*, in Compatible spatial discretizations, Springer, 2006, pp. 89–119.
- [5] M. CALVO AND E. HAIRER, *Accurate long-term integration of dynamical systems*, Appl. Numer. Math., 18 (1995), pp. 95–105.
- [6] M. CALVO AND J. SANZ-SERNA, *The development of variable-step symplectic integrators, with application to the two-body problem*, SIAM J. Sci. Comput., 14 (1993), pp. 936–952.
- [7] E. CELLEDONI, V. GRIMM, R. MCLACHLAN, D. MCLAREN, D. O’NEALE, B. OWREN, AND G. QUISPTEL, *Preserving energy resp. dissipation in numerical PDEs using the average vector field method*, Journal of Computational Physics, 231 (2012), pp. 6770–6789.
- [8] S. H. CHRISTIANSEN, H. MUNTKE-KAAS, AND B. OWREN, *Topics in structure-preserving discretization*, Acta Numerica, 20 (2011), pp. 1–119.
- [9] R. COURANT, K. FRIEDRICH, AND H. LEWY, *On the Partial Difference Equations of Mathematical Physics*, IBM Journal of Research and Development, 11 (1967), p. 215.
- [10] M. DAHLBY AND B. OWREN, *A General Framework for Deriving Integral Preserving Numerical Methods for PDEs*, SIAM J. Sci. Comput., 33 (2011), pp. 2318–2340.
- [11] M. DAHLBY, B. OWREN, AND T. YAGUCHI, *Preserving multiple first integrals by discrete gradients*, J. Phys. A: Math. Theor., 44 (2011).
- [12] D. FURIHATA AND T. MATSUO, *Discrete Variational Derivative Method: A Structure-Preserving Numerical Method for Partial Differential Equations*, CRC Press, 2010.
- [13] O. GONZALEZ, *Time integration and discrete Hamiltonian systems*, J. Nonlinear Science, 6 (1996), pp. 449–467.
- [14] E. HAIRER, C. LUBICH, AND G. WANNER, *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, Springer, Berlin, 2006.
- [15] A. N. HIRANI, *Discrete exterior calculus*, PhD thesis, California Institute of Technology, 2003.
- [16] R. A. LABUDDE AND D. GREENSPAN, *Energy and momentum conserving methods of arbitrary order for the numerical integration of equations of motion*, Numerische Mathematik, 25 (1975), pp. 323–346.
- [17] B. LEIMKUHLER AND S. REICH, *Simulating Hamiltonian dynamics*, Cambridge University Press, Cambridge, 2004.
- [18] S. LI AND L. VU-QUOC, *Finite difference calculus invariant structure of a class of algorithms for the nonlinear Klein-Gordon equation*, SIAM J. Numer. Anal., 32 (1995), pp. 1839–1875.
- [19] J. E. MARSDEN AND M. WEST, *Discrete Mechanics and Variational Integrators*, Acta Numerica, (2001), pp. 1–158.
- [20] J. MUNKRES, *Topology*, Prentice Hall, 2 ed., 2000.
- [21] P. J. OLVER, *Applications of Lie Groups to Differential Equations*, Springer, 2nd ed., 2000.
- [22] P. J. OLVER, *Geometric foundations of numerical algorithms and symmetry*, Appl. Alg. Engin. Comp. Commun., 11 (2001), pp. 417–436.
- [23] G. QUISPTEL AND D. MCLAREN, *A new class of energy-preserving numerical integration methods*, J. Phys. A: Math. Theor., 41 (2008).
- [24] J. SIMO, N. TARNOW, AND K. WONG, *Exact energy-momentum conserving algorithms and symplectic schemes for nonlinear dynamics*, Computer Methods in Applied Mechanics and Engineering, 100 (1992), pp. 63–116.
- [25] A. T. S. WAN, A. BIHLO, AND J.-C. NAVE, *The Multiplier Method to Construct Conservative Finite Difference Schemes for Ordinary and Partial Differential Equations*, SIAM J. Numer. Anal., 54 (2016), pp. 86–119.